Results of the 2006 Spoken Term Detection Evaluation

Jonathan G. Fiscus, Jerome Ajot, John S. Garofolo
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
{jfiscus,ajot,jgarofolo}@nist.gov

George Doddingtion
Orinda, CA, USA
george.doddington@comcast.net

ABSTRACT

This paper presents the pilot evaluation of Spoken Term Detection technologies, held during the latter part of 2006. Spoken Term Detection systems rapidly detect the presence of a *term*, which is a sequence of words consecutively spoken, in a large audio corpus of heterogeneous speech material. The paper describes the evaluation task posed to Spoken Term Detection systems, the evaluation methodologies, the Arabic, English and Mandarin evaluation corpora, and the results of the evaluation. Ten participants submitted systems for the evaluation.

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval

General Terms

Measurement, Design, Theory.

Keywords

Speech Retrieval, Audio Indexing, Audio Mining, Multilingual, Speech Recognition

1. INTRODUCTION

Information processing has become a major activity in the world, and spoken communications is a major source of that information. This, coupled with growing computer-accessible volumes of audio data, has created an opportunity and a need for effective retrieval of information from archives of speech data. To support development of such technology, NIST created the Spoken Term Detection (STD) pilot evaluation initiative. This evaluation is structured as a collaborative research activity that is intended to foster technical progress in STD, with the goals of exploring promising new ideas in STD, developing advanced technology incorporating these ideas, measuring the performance of this technology, and establishing a community for the exchange of research results and technical insights. The evaluation supported experiments on three languages: Arabic (Modern Standard and Levantine), English, and Mandarin Chinese.

The evaluation task and evaluation infrastructure are documented in the STD 2006 Evaluation Plan which can be found on the NIST STD website [1]. Section 1 summarizes the evaluation plan which defines: the STD task and STD system architecture, the STD system output, the STD search terms, and the evaluation methodology. Section 2 covers the specifics of the STD 2006 evaluation including the test corpora and results.

1.1. STD Task and System Architecture

The goal of the STD evaluation task is to rapidly detect the presence of a *term* – a sequence of words consecutively spoken – in a large audio corpus of heterogeneous speech material. The effectiveness of a deployed STD system is a tradeoff between processing resource requirements and detection accuracy. The evaluation plan prescribes a generic system architecture (Figure 1) that systems must adhere to in order to participate in the evaluation. While NIST typically does not prescribe system-internal operations for its language technology evaluations, it was necessary to model two key application constraints so that the evaluation task was a good model of the intended application. First, search times for a given term must be small (within seconds). Therefore, systems must index the audio corpus before searching, rather than search the corpus directly for each search term. Second, the indexer does not have advance knowledge of the search terms and therefore cannot use that information during indexing. These imposed constraints effectively force system developers to address both the real-time challenge of pre indexing corpora without knowledge of the search terms and the challenge of rapidly returning search results.

A benefit of the prescribed architecture is to enable uniform operation resource measurements across systems, e.g., indexing speed, index size, search speed, etc.

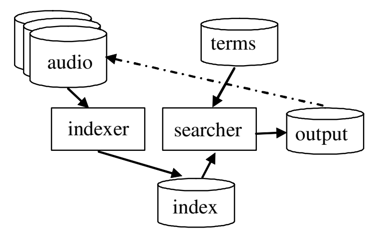

Figure 1: Generic STD System Architecture

Previous speech retrieval evaluations like TREC's Spoken Document Retrieval [7] (SDR), and Topic Detection and Tracking [8] (TDT) have investigated technologies similar to STD. However, they each addressed different problems. Source data robustness is a key component of STD whereas SDR and TDT focused on the broadcast news domain. The query for STD, a search term, is a markedly smaller unit than SDR's query definition which was a natural language description of an information need, and more specific than TDT's topic exemplar documents. A technology similar to STD is keyword spotting [10]. The main difference between keyword spotting and STD is the number of words in a search term.

1.2. STD Terms

STD terms are sequences of consecutively spoken words. They have no linguistically defined correlate, but range in grammatical scope from single words to phrases, e.g, /grasshopper/, /organizing/, /New York/, /Albert Einstein/, and /the coalition government/. STD terms are required to have a "recognizable, complete meaning" that a hypothetical user would want to find. For example, the trigram /crosby v. o./ is not a potential term because the /crosby/ is the name of a Voice Of America (/v. o. a./) reporter /Tom Crosby/ and therefore not complete. Further, it is not recognizable by itself. While this is a subjective definition, it models the information need of the searcher.

Native language orthography is the sole specification of a search term. Defining terms in this manner was a pragmatic decision. Ideally, each term would have a single specific interpretation or meaning. However, the contextual/phonetic definitions required to differentiate senses is beyond the term specification a hypothetical user will perform. Therefore, the term definitions for /wind/ (air movement) and /wind/ (twist) are indistinguishable. Systems must therefore handle pronunciation variations internally.

Terms include five or fewer "words." The concept of a "word" is not the same in all languages. In English, words include the morphological prefixes and suffixes in typical written text. Since articles, pronouns, and prepositions are separate words in English, they were not included. In Arabic, words are declared to be white space separated elements as typically used in Modern Standard Arabic (MSA). The Arabic terms included particles as part of the term since they are affixes and prefixes. For Mandarin, word segmentation was a product of the transcription process where the transcribers divided the character streams into word-like units.

Human annotators selected terms for the evaluation from a series of putative term lists derived from the evaluation corpus and from out-of-corpus sources. The in-corpus putative term lists included: tri-grams, bi-grams, uni-grams, and high frequency words. Bi-grams of all selected tri-grams and uni-grams of all bi-gram terms (including the bi-grams of the selected tri-grams) were added to the term lists so that constituent error rates for multi-word terms could be measured. Annotators added terms to the term lists that did not occur in the evaluation corpus. These out-of-corpus terms were used test the system's response to non-occurring terms.

Reference term occurrences are found automatically by searching high-quality transcripts. The following criteria were employed to determine the existence of a term; constituent words of a term must be adjacent, spoken by a single speaker, and within 0.5 second of each other. Sub-strings were not considered matches so an uttered word /grasshopper/ was not an occurrence of the term /grass/. Likewise, inflected forms were not considered matches so an uttered word /speaking/ was not an occurrence of the term /speak/. In a real applications, these forms could be sought simultaneously if that is what the user wishes.

1.3. STD System Inputs and Outputs

The ability of STD systems to process a variety of sources is an important factor of system performance, so the evaluation corpus contains as many sources as possible. STD systems index and search the complete test corpus with no *a priori* knowledge of the data. However, to make the first evaluation tractable for simple ports of existing technology, the audio files within the evaluation corpus included domain identifications, e.g., broadcast news (BNEWS), conversational telephone speech (CTS), or meeting room (MTG). Future STD evaluations will not provide this side information.

Systems process each term independently during the system's search phase. For each likely occurrence of a given term, the system is required to output a record that includes:

- the beginning and ending time of the term occurrence in the audio recording.
- a binary decision ("YES" or "NO") as to whether or not the system believes this putative occurrence is an occurrence of the term. This is called an "actual decision." Internal to the system, an actual decision threshold differentiates the YES/NO decisions[1].
- a detection score indicating how likely this putative term actually occurs (with more positive values indicating more likely occurrences.) The score for each term occurrence can be of any scale. However, the scores must be on a commensurate scale to permit the generation of pooled-term performance measurements.

Requiring systems to output both an actual decision and detection score for each putative term occurrence has a large benefit for system evaluation. Developers need a single metric to optimize system performance. However, *a priori* specification of an optimization criterion is dependent on the application: i.e. is high precision or high recall required. The actual decision provides the means to both optimize performance to a specific optimization criterion, via "YES" actual decisions, and over-generate putative occurrences, via "NO" actual decisions, to assess performance over a wide range of operating points. Section 1.4 covers this in more detail.

1.4. STD Evaluation Methodology

STD is a detection task – namely to detect all of the occurrences of each given term in the audio corpus. Two error types characterize STD performance: false alarms and missed detections.

Several NIST language evaluations have used the detection evaluation formalism, e.g., as in speaker recognition [1] [3]. Abstractly, detection systems answer the question: "Is this instance of data an example of the provided training data?" Each time the system answers this question, it is a called a "trial". The instance can be anything, a segment of speech for instance. The training data can be an exemplar of any form, a set of speech files for instance. Typically, the instances are discrete events or objects and therefore the trials are discrete. However, the STD task lacks the usual structure of discrete 'trials' necessary for computing normalized error rates, and therefore the evaluation methodology was adapted as follows.

---

[1] System performance is optimized by computing system performance based on the actual decisions.

- First, an estimate was required for the number of discrete trials in the reference. Unlike the speaker recognition evaluations, there are no discrete trials in continuous speech. Thus, part of the evaluation metric below specifies the number of trials as a constant.
- Second, an alignment between the system-detected occurrences and reference occurrences was needed in order to evaluate the system because systems are not given *a priori* knowledge of word/term boundaries in the speech. The Hungarian Solution to the Bipartite Graph [9] matching problem was used to compute the 1:1 mapping. The optimized objective function takes into account the temporal overlap of the system and reference occurrences (with a tolerance collar) and the term occurrence's detection score.
- Third, systems generate only a partial list of putative term occurrences[1] unlike speaker evaluations where systems provided decisions and scores for every trial.

System performance was evaluated using two methods: graphically with Detection Error Tradeoff (DET) curves [3] and for a particular operating point in the DET curve space using a Term-Weighted Value (TWV). The former provides an intuitive view of system performance for both high recall and high precision application needs, while the TWV provides developers with a single performance metric as a target for system optimization.

1.4.1. Detection Error Tradeoff Curves

Graphical performance assessment uses a detection error tradeoff (DET) curve that plots miss probability (PMiss) versus false alarm probability (PFA). Miss and false alarm probabilities are functions of the detection threshold, θ. This (θ) is applied to the system's detection scores, which are computed separately for each search term, then averaged to generate a DET line trace. The formulas for a single term's PMiss and PFA are:

$$P_{\mathrm{Miss}}(term,\theta) = 1 - \frac{N_{\mathrm{correct}}(term,\theta)}{N_{\mathrm{true}}(term)}$$

$$P_{\mathrm{FA}}(term,\theta) = \frac{N_{\mathrm{spurious}}(term,\theta)}{N_{\mathrm{NT}}(term)}$$

where:

$N_{correct}(term,\theta)$ is the number of correct (true) detections of *term* with a detection score greater than or equal to θ.

$N_{spurious}(term,\theta)$ is the number of spurious (incorrect) detections of term with a detection score greater than or equal to θ.

N_{true}(term) is the true number of occurrences of term in the corpus,

N_{NT}(term) is the number of opportunities for incorrect detection of term in the corpus (= "Non-Target" *term* trials).

[1] The general application would also preclude generating exhaustive putative occurrences.

Since there is no discrete specification of "trials", the number of Non-Target trials for a term, $N_{NT}(term)$, is defined somewhat arbitrarily to be proportional to the number of seconds of speech in the test set. Specifically:

$$N_{\mathrm{NT}}(term) = n_{\mathrm{tps}} \cdot T_{\mathrm{speech}} - N_{\mathrm{true}}(term)$$

where:

n_{tps} is the number of trials per second of speech (arbitrarily set to 1), and

T_{speech} is the total amount of speech in the test data (in seconds).

1.4.2. Term Weighted Value

To measure a system's "value" is to measure the usefulness of a system to a user. A perfect system always responds correctly to a stimulus, however an omitted response or a misleading response reduces the value of a system to a user. Thus, Term-Weighted Value (TWV) is one minus the average value lost by the system per term. The value lost by the system is a weighted linear combination of P_{Miss} and P_{FA} as defined above. The weight, β, takes into account both the prior probability of a term and the relative weights for each error type.

$$\mathrm{TWV}(\theta) = 1 - \underset{term}{average}\left\{P_{\mathrm{Miss}}(term,\theta) + \beta \cdot P_{\mathrm{FA}}(term,\theta)\right\}$$

where:

$$\beta = \frac{\mathbf{C}}{\mathbf{V}} \cdot \left(\mathbf{Pr}_{\mathbf{term}}^{-1} - 1\right).$$

θ is the detection threshold.

For the current evaluation, the cost/value ratio, **C/V**, is 0.1, thus the value lost by a false alarm is a tenth of the value lost for a miss. The prior probability of a term, $\mathbf{Pr}_{\mathbf{term}}$, is 10^{-4}.

The maximum possible TWV is 1.0, corresponding to "perfect" system output: no misses and no false alarms. The TWV of a system that outputs nothing is 0.0 and negative TWVs are possible.

1.4.3. Actual vs. Maximum Term Weighted Value

While DET curves represent performance for all possible values of θ, two points on the DET curve are of interest because they determine if the system's actual decision threshold is optimal. The first is Actual Term-Weighted Value (ATWV) which is the TWV using the actual decisions. ATWV represents the system's ability to predict the optimal operating point given the TWV scoring metric. The second is Maximum Term-Weighted Value (MTWV). MTWV is the TWV at the point on the DET curve where a value of θ yields the maximum TWV. The difference between the values for ATWV and MTWV indicate the benefit of selecting a better actual decision threshold.

1.5. Processing Resource Measurements

Fielded STD technologies will process vast amounts of data. As such, "speed is important". Systems were required to record speed and resource measurements during processing. The measurements allow both extrapolations to larger data sets and facilitate inter-system comparisons, i.e., comparing fast to slow systems would be unfair. The measurements are Index Size, Indexing Speed, Indexing Memory Usage, Search Speed, and Search Memory Usage.

Measurements such as these are often difficult to make during system execution when the processes are broken down into sub steps via UNIX shell scripts, (which the researchers predominately use.) To facilitate the measurements, NIST developed a new tool, ProcGraph [11], that tracks resource usage for UNIX shell scripts including subordinate processes.

2. STD 2006 EVALUATION

The 2006 evaluation was the first STD Evaluation. The process of designing the evaluation began in spring 2006. During the summer and fall of 2006, NIST assembled the evaluation infrastructure and developers built their systems. The evaluation occurred in November. NIST hosted the 2006 STD Evaluation workshop to discuss the results of the evaluation on December 14-15, 2006.

Ten sites participated in the evaluation: BBN Technologies (BBN), Brno Univ. of Tech. (BUT), Department of Defense (DOD), IBM, Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), OGI School of Science and Tech. (OGI), Queensland Univ. of Tech. (QUT), SRI International (SRI), Stellenbosch Univ. (STELL), Technischen Universität Berlin (TUB). STELL and TUB collaborated to submit a system referred to a STBU.

The following sections provide summaries of the evaluation corpora, terms, and system performance measurements.

2.1. Evaluation Corpora

The evaluation made use of a small corpus of previously used Speech-To-Text evaluation test sets [4], [5], [6] which included high quality transcripts and automatically-derived time locations for each word. The word locations where computed with two methods. The first method, which was used for the English data, made use of Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) speech recognition tools to align the reference transcript to the acoustic signal. (Forced word alignment is the common name for this process.) The second method, which was used for the Arabic and Mandarin data, inferred word locations from the output of an automatic speech recognizer (ASR) by finding a word alignment between the reference and ASR output words, then mapping the ASR word times onto the reference words. The time mapping procedure linearly interpolated times for reference words during regions of incorrectly recognized speech. As expected, they were not as accurate as the forced alignment-derived word times; however, the use of a temporal mapping tolerance collar reduces the impact of less accurate word times. Table 1 lists the data for each language and source type with the predominant dialect identified.

The Linguistic Data Consortium[1] transcribed all if the material according to high quality standards set by the speech recognition community. Appen[2] further enhanced the Arabic transcripts by correcting minor flaws and adding diacritics to the transcripts.

Table 1: STD 2006 evaluation corpus composition

	Arabic	Chinese	English
Broadcast News (BNEWS)	MSA ~1 hour	Mandarin ~1 hour	American~ 3 hours
Telephone Conversations (CTS)	Levantine ~1 hour	Mandarin ~1 hour	American ~3 hours
Roundtable Meetings (MTG)	None	None	American ~2 hours

2.2. Evaluation Terms

Nominally, 1100 terms were selected for each language with the following rough proportions: 10% tri-grams, 40% bi-grams, 50% uni-grams. For the Arabic data, the vast majority tri-grams were partial sentences and whole sentences. Since they were linguistically larger than phrases, Arabic tri-grams were not included in the term lists.

Table 2 shows the number of terms selected per language and the number of reference occurrences per source type. The terms selection protocol produced an English term list balanced by source type. However, the same is not true for Arabic and Mandarin. Subsequent evaluations will factor source type into the term selection protocol.

The evaluation used two forms of the Arabic terms, with and without diacritics – the former being posited as a means to better specify the terms thus accounting for dialectal variation. The diacritized terms were derived from the non-diacritized terms by a process that converted each term into a set of diacritized variants. The diacritized variants for each constituent word were limited to the variations found in the reference transcripts.

Table 2: Term Set Properties by Language

	Arabic		English	Mandarin
	Diacritized	Non-Diacritized		
Terms Selected	1101	937	1100	1120
Ref. Occ.	2433	2807	14421	3684
Reference Occurrences Per Source, Per Speech Hour				
BNEWS	1513	1749	2212	3070
CTS	557	638	1957	582
MTG			1750	

2.3. Arabic Results

BBN, BUT and DOD participated in the Arabic test. Table 3 summarizes their scores for both diacritized and non-diacritized terms. The highest ATWV for non-diacritized terms in the CTS domain was 0.34 by BBN. For the diacritized terms in the BNEWS domain, the highest ATWV was –0.06.

[1] See the LDC website www.ldc.upenn.edu

[2] See the Appen website www.appen.com.au

Although we wanted to test our hypothesis that diacritics would help searching, two difficulties emerged during the evaluation that prevented us from doing so. First, and foremost, diacritization is an inherently difficulty task for humans and therefore the reference transcripts contained diacritization errors. For example, Appen had two independent teams correct and diacritize 25 minutes of BNEWS and CTS data (50 minutes total). 12.5% of the BNEWS and 11% CTS words had at least one different diacritic after quality control passes. To put this in context, this is two-three times the error rate of human transcription of English. Second, building purely undiacritized systems is not possible because common Arabic transcription practices make use of diacritics to disambiguate word usage. Thus, the evaluation results between the two term sets are not directly comparable.

Table 3: Arabic Actual Term Weighted Values

Search Terms	Site	BNEWS	CTS
Non-diacritized	BBN		0.35
Diacritized	BUT	-0.09	0.00
	DOD		-6.57

2.4. English Results

All sites built systems for the English data. (BBN and DOD only built systems for the CTS portion of the test set.) Figure 2 presents the ATWVs for all the English tests by source type. The highest ATWVs were 0.85 for BBN's system on BNEWS data, 0.83 for BBN's system on CTS data, and 0.26 for SRI's system on MTG data. As expected, the order of difficulty by source type is BNEWS, CTS, MTG. This matches the source difficulty for speech recognition systems in the Rich Transcription evaluations.

Figure 4 contains the DET curves for all primary English systems on the CTS data. The graph shows the tradeoff between false alarms and missed detections. DET line traces for better performing systems, with regard to accuracy, have lines closer to the origin. The BBN system, which had the highest ATWV at 0.83, achieved a MTWV of 0.83 indicating a suitable actual decision threshold was chosen. At the MTW point, the false alarm rate was 0.005% and a missed detection rate was 11.9%. Note that no DET curve trace extends beyond 5% miss because the systems do not output a decision for every trial.

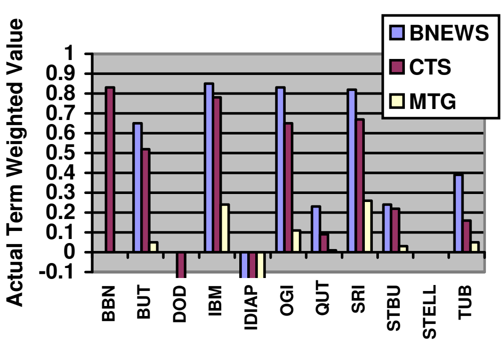

Figure 2: English Actual Term Weighted Values

The high ATWVs indicate developers strove to build accurate systems. For this initial evaluation though, most developers did not have the resources to build fast systems. Instead, developers used existing language technologies to build their STD systems. Figure 3 shows the performance of systems as a function of Indexing Speed measured in processing hours per indexed speech hours. On this graph, scores that appear in the upper left quadrant are better because they indicate accurate and fast STD systems. With the wide range of indexing speeds, it would be difficult to quantify the tradeoff with a single measurement that combines accuracy and speed into a single measure. Instead, next year's evaluation will likely require specific processing speed thresholds (e.g., 0.01, 0.1, and 1.0 Indexing Speeds) so that processing speed can be controlled while accuracy is measured.

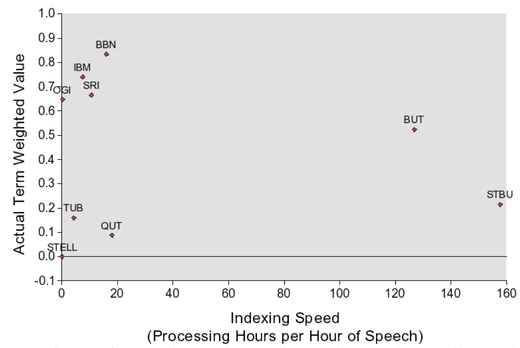

Figure 3: ATWV as a Function of Indexing Speed for the CTS data

2.5. Mandarin Results

BBN and DOD participated in the Mandarin tests and achieved scores of 0.38 and –1.02 ATWV respectively on the CTS data set. Neither participant processed the BNEWS data.

The evaluation infrastructure relied on human segmentation for both term selection and reference term location. This was acceptable for term selection because a human was in the loop. However, we are studying whether or not word segmentation negatively affected the scoring.

3. CONCLUSIONS

NIST conducted a pilot evaluation of Spoken Term Detection systems in December 2006. The evaluation was successful in that: it drew a significant number of participants (10) for a first such evaluation; the evaluation proved the feasibility of the STD technology measurement approach; it provided a useful baseline for future work; it touched on challenges with regard to technology robustness including speed, scalability, multilinguality, and domain independence. While the challenges of scalability and domain independence were not fully explored in the pilot, the evaluation set the stage for future efforts which explore these important dimensions in more depth.

The evaluation resulted in all ten of the participants having developed systems to process the English Conversational Telephone Speech subset of the test data. The highest ATWV for these systems was 0.83. The indexing speeds for these systems were extremely variable -- ranging from 0.168 to 157.6 processing-hours-per-hour-of-speech in the test corpus.

The most important advance to measurement science from this effort was the adaptation of the detection evaluation methodology to STD. In the course of creating the metric for this task, we developed a new approach which permitted us to measure detection accuracy when the events to be detected are not discrete trials.

Furthermore, the evaluation components developed to map system-to-reference term occurrences and build partial DET curves will be useful for a variety of other detection-oriented evaluations.

We intend to expand the scope of future STD evaluations to address the scalability and domain diversity issues and we will continue to study and refine the evaluation protocol with regard to: a term selection process that exercises the depth and breadth of the application domain in the most effective and informative manner, an assessment of the impact of transcription accuracy on performance measurements, develop metrics that combine accuracy and speed in informative and intuitive ways, improve the consistency of Arabic term diacritization, and assess the impact of Mandarin word segmentation. Toward this end, we expect to run a second STD evaluation in 2008 using a much larger and more diverse test set. The evaluation will challenge the technology in two dimensions: data robustness and processing speeds. The evaluation data will include a wider variety of data and processing speed will play a major role in the evaluation of systems.

4. DISCLAIMER

These tests are designed for local implementation by each participant. The reported results are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U. S. Government. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

5. REFERENCES

[1] "2006 STD Website and Evaluation Plan", http://www.nist.gov/speech/tests/std/std2006/.

[2] Przybocki, M., Martin, A., "NIST Speaker Recognition Evaluation Chronicles", Proceedings of Odyssey 2004.

[3] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., "The DET Curve in Assessment of Detection Task Performance", Eurospeech 1997, pp 1895-1898.

[4] Fiscus et. al., "Results of the Fall 2004 STT and MDE Evaluation", RT-04F Evaluation Workshop Proceedings, November 7-10, 2004.

[5] Fiscus, J., et al., "The Rich Transcription 2005 Spring Recognition Meeting Evaluation", 2nd International Workshop on Machine Learning for Multimodal Interaction, LNCS 3869.

[6] Fiscus, J., Ajot, J., Michel, M., Garofolo, J., "The Rich Transcription 2006 Spring Meeting Recognition Evaluation", 3rd International Workshop on Machine Learning for Multimodal Interaction, LNCS 4299.

[7] Garofolo, J., Auzanne, C., Vorhees, E., "The TREC Spoken Document Retrieval Track : A Success Story", Proceedings of the Recherche d'Informations Assiste par Ordinateur: Content Based Multimedia Information Access Conference, April 12-14, 2000

[8] Allan, J., "Topic Detection and Tracking: Event-based Information Organization", ISBN 978-0792376644

[9] Harold W. Kuhn, "The Hungarian Method for the assignment problem", *Naval Research Logistic Quarterly*, **2**:83-97, 1955.

[10] Rose, R. C., Paul, D. B., "A Hidden Markov Model Based Keyword Recognition System", 1990 International Conference on Acoustics, Speech, and Signal Processing, 1990, pp. 129-132 vol.1.

[11] http://www.nist.gov/speech/tools/index.htm

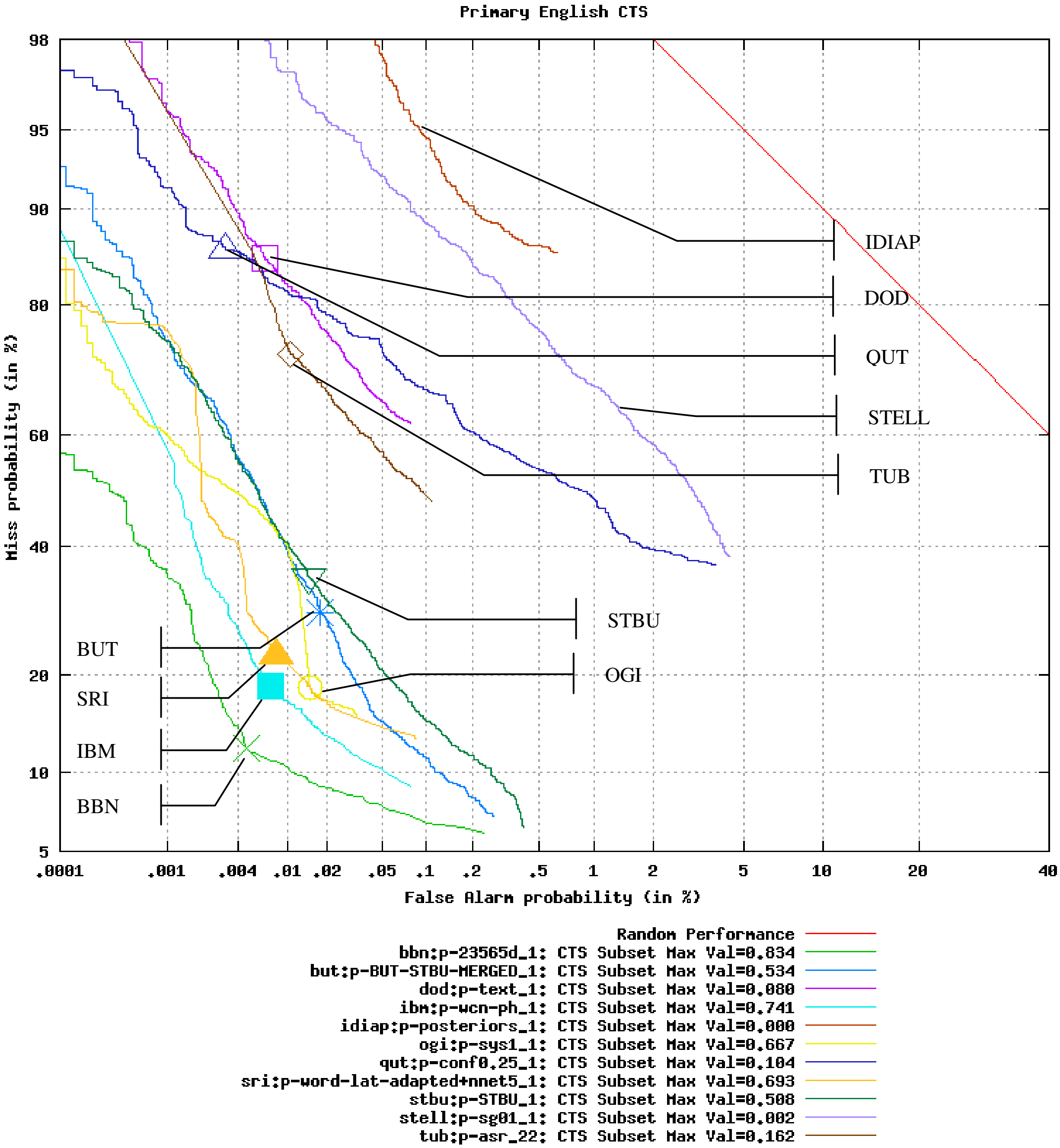

Figure 4: DET Curve for English, CTS Primary Systems. The symbols on the chart is the point of Maximum ATWV